DOCUMENT RESUME

ED 262 050 TM 850 509

AUTHOR Phillips, Gary W.
TITLE Latent Trait Approach to Domain Score Estimation.
PUB DATE Mar 82
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; Estimation (Mathematics); Goodness of Fit; *Item Analysis; *Latent Trait Theory; Mastery Tests; *Mathematical Models; Scores; Statistical Analysis; *Test Construction; Test Items; Test Theory
IDENTIFIERS Binomial Error Model; Classical Test Theory; *Domain Referenced Tests

ABSTRACT

This paper presents an introduction to the use of latent trait models for the estimation of domain scores. It was shown that these models provided an advantage over classical test theory and binomial error models in that unbiased estimates of true domain scores could be obtained even when items were not randomly selected from a universe of items. This allowed the full use of item statistics during test development. Furthermore, it was argued that when all domain subtests measured the same underlying trait, then all subtest domain scales might be related to each other via the underlying theta scale. Finally, a method was presented whereby the unidimensionality assumption could be tentatively tested. This was to fit a LISREL congeneric measurement model to the observed subtest variance-covariance matrix. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262050

LATENT TRAIT APPROACH TO DOMAIN SCORE ESTIMATION

Gary W. Phillips

Maryland State Department of Education

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the annual meeting of the American Educational Research Association, New York, March 1982.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Phillips, Gary W.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 509

LATENT TRAIT APPROACH TO DOMAIN SCORE ESTIMATION

The most commonly used score in criterion-referenced testing is the domain score. The domain score is radically different from norm-referenced scores such as the percentile rank. This latter score is related to the proportion of examinees below a given mastery level in a well-defined population, whereas domain scores refer to the proportion of mastery in a well-defined content area. This paper reviews three psychometric conceptualizations of domain scores. These are classical test theory, the binomial error model, and latent trait models. The first two approaches will only be briefly covered with the third the subject of the bulk of the paper. Although there are other ways of viewing domain scores, these three models were chosen because they are so interrelated. In addition, the first two models have been the most widely used methods of dealing with domain scores.

The paper will argue that latent trait models provide several advantages over the classical test theory and binomial error models. The major advantage discussed here is that they provide a means whereby unbiased estimates of true domain scores can be obtained even though items have been non-randomly sampled from a population of items. Because of the need to randomly sample items, many researchers in criterion-referenced

testing have argued against reliance on item statistics during test development. For example, Millman states "Should item difficulty and discrimination indexes be used to choose items? No! Selection of items on these criteria can result in a test where the items are not representative of the domain in difficulty level or in the underlying attributes being measured" (Millman, 1974, p. 339).

This paper will show that this is not true within the context of latent trait models. Using these models the test developer can make full use of item statistics, non-randomly selected items (e.g., for test targeting purposes) and still obtain unbiased estimates of true domain scores.

Classical Test Theory

In this paper a domain score is defined as the proportion of items answered correctly on an N-item test (i.e., $Z=X/N$, where X is the raw score). The classical test theory approach assumes that the observed domain score Z is the sum of a true domain score ζ and an error item e. In this conceptualization the observed domain score is an unbiased estimate of the true domain score ζ if the items have been randomly and representatively sampled from a well-defined population of items. In the classical test theory model the squared standard error of measurement is

$$\sigma_e^2=\sigma_z^2-\sigma_\zeta^2$$

One problem with this model for domain scores is that the estimate of ζ is unstable when the number of items is small. A more stable estimate of ζ can be obtained by taking into account the other group scores (Lord and Novick, 1968, p. 65). Using this approach the best estimate of ζ becomes

$$\hat{\zeta}=\rho(X/N)+(1-\rho)(\mu/N),$$

where ρ and μ equal the classical test theory reliability coefficient and the mean of the examinee group respectively.

There have been many fruitful elaborations of the classical test theory model. For example, Livingston (1972) adapted the model to the case where deviations are taken from a passing

score rather than the mean of the domain scale. Brennan (1980) extended this by conceptualizing domain scores within the context of generalizability theory (Brennan, 1980). Others have used Bayesian estimation methods to improve the domain score estimates by incorporating prior information (for a review of this literature, see Hambleton, et al, 1978).

Although classical test theory has been used extensively as a psychometric model for domain scores, it has several serious shortcomings. The major problem with the model is that examinee and test characteristics are inseparable. Consequently, the measurement error and reliability of the domain scale is affected by examinee variability.

Furthermore, classical test theory assumes that the standard error of measurement and test reliability are the same throughout the domain scale. This is an undesirable characteristic as the test may clearly provide better measurements in some intervals along the scale. Related to this is the fact that the distribution of error is usually assumed to be normal throughout the scale. Again, this assumption is unrealistic especially for near zero or perfect true scores.

Finally, the model requires the random selection of items from a population of items. As mentioned above, since the selection of items must be random, the role of item statistics is usually downplayed by users of classical test theory models.

Binomial Error Model

For tests of reasonable length the binomial error model has been used as a psychometric approach to domain scores. In this model it is assumed that the observed domain score Z is obtained from a random sample of N equally difficult, dichotomously scored items. This is identical to assuming that the examinees' response to each item is a Bernoulli trial with the probability of success equal to ζ. Under these assumptions, the examinees' observed raw score on the test is $X=N\zeta+Ne$, and the conditional relative frequency distribution of X given ζ on an N item test is

$$f(X|\zeta)=\binom{N}{X}\zeta^{X}(1-\zeta)^{N-X},$$

for X=0, 1, .., N, and where

$$\binom{N}{X}=\frac{N!}{X!(N-X)!}\ .$$

In the above familiar binomial distribution $\binom{N}{X}$ is the number of combinations (or patterns) of N items that can result in a raw score X. The quantity $\zeta^{X}(1-\zeta)^{N-X}$ is the probability of obtaining one of the combinations. The true domain score ζ is conceptualized as the examinees' true ability on the test.

According to this model the squared standard error of measurement of X and Z are

$$\sigma^2_{X|N\zeta}=N\zeta(1-\zeta), \text{ and}$$

$$\sigma^2_{Z|\zeta}=\frac{1}{N^2}[N\zeta(1-\zeta)], \text{ respectively.}$$

The squared standard error of measurement is maximum when ζ=.50 and approaches zero as ζ approaches 0 or 1. The relationship between the squared standard error of measurement of classical test theory and the binomial error model is

$$\sigma_e^2 = \sum_{i=1}^{n} \sigma_{z|\zeta}^2 / n$$

where n equals the number of examinees. The above equation shows that σ_e^2 is the average $\sigma_{z|\zeta}^2$ in a population of examinees.

This model has been very useful in solving different criterion-referenced problems. For example, Millman (1972) used the model to generate a set of tables that indicate the error rates associated with various true scores, test lengths, and passing scores. A further development of the model assumes that the distribution of ζ takes the form of a two-parameter beta variable (Lord, 1968, p. 520). Using this beta-binomial approach various investigators have provided solutions to problems of test length (Novick and Lewis, 1974). passing scores (Huynh, 1977), and reliability (Huynh, 1976, 1977).

The binomial error model has one characteristic which makes it desirable as a psychometric model for domain scores. This is that it provides estimates of measurement error which vary across the domain scale and do not depend on examinee ability distributions.

Unfortunately, the binomial error model has two disadvantages which make it less useful than the latent trait approach to be mentioned next. First, the model assumes that all items are equally difficult. This assumption is seldom even approximated in practical testing situations. Secondly, like classical test theory, the model requires that items be randomly sampled from a population of items.

Latent Trait Models

The remainder of this paper will explore the use of logistic latent trait models as a means of estimating and characterizing the behavior of domain scores. In logistic latent trait models the probability of a correct response to an item is described by a one-, two-, or three-parameter item response function called an item characteristic curve (ICC). In this paper it will be assumed that ICCs can be sufficiently described by the one-parameter (Rasch) model. However, all of the conclusions are equally applicable to the more general two-, and three-parameter models.

Under the one-parameter model it is assumed that items differ from each other only in difficulty. When this is the case it is known that the conditional relative frequency distribution of X given θ (where θ is the underlying unidimensional latent trait being measured by each item) is a compound binomial distribution (Lord and Novick, 1968, p. 525). When raw scores

X are transformed to domain scores Z the mean and variance of the conditional relative frequency distribution of Z given θ are

$$\mu_{z|\theta}=\zeta(\theta)=\frac{1}{N}\sum_{i=1}^{N} P_i(\theta), \text{ and}$$

$$\sigma^2_{z|\theta}=\frac{1}{N^2}\sum_{i=1}^{N} P_i(\theta)Q_i(\theta), \text{ respectively.}$$

In the first of the above equations P_i and Q_i equal the probability of correctly and incorrectly answering the ith item at θ and ζ is the true domain score resulting from the regression of Z on θ and is called the test characteristic curve (TCC). For the Z scale, it is equal to the mean ICC on an N item test. In the second equation $\sigma^2_{z|\theta}$ is the squared standard error of measurement for the domain scale. The relation between $\sigma^2_{z|\theta}$ and the classical test theory squared standard error of measurement σ^2_e is

$$\sigma^2_e=\frac{1}{n}\sum_{j=1}^{n}\sigma^2_{z|\theta},$$

for j=1, 2, .., n, where n=the number of examinees. From the above equation it can be seen that the squared standard error of measurement of classical test theory equals the average $\sigma^2_{z|\theta}$ across n examinees. Furthermore, in the restrictive case when all items have identical ICCs (i.e., all items are equally difficult) the $\sigma^2_{z|\theta}$ equals the squared standard error of measurement of the binomial error model $\sigma^2_{z|\zeta}$.

Latent trait models provide several advantages over classical test theory and binomial error models. Like the binomial error model, latent trait models provide different standard errors of measurement throughout the domain scale that are independent of examinee ability distributions. However, unlike the binomial error model, it is not necessary to assume equally difficult items. In fact, with the two- and three-parameter latent trait models item discrimination and guessing also may be assumed to vary across items.

The second advantage of latent trait models is that items can be non-randomly sampled, yet, unbiased estimates of ζ can still be obtained. This characteristic of the model allows test developers to make full use of item statistics in the construction of a test. For example, if the domain scale has a passing standard then the test developer may wish to build maximum precision into the scale at the cut score by choosing items with minimum measurement error around the cut score. This type of test targeting strategy is not possible with classical test theory and binomial error models.

Figure 1 shows an example of a latent trait approach to estimating domain scores. In Figure 1, two domain scale TCCs have been graphed. Let's assume the TCC on the left is based on the one-parameter calibrations of all items in an item bank. Furthermore, let's assume that each item in the bank has been written with an explicit set of item specifications and

that the items in the bank represent a random sample from a universe of items that could have been written from the same specifications. Under these conditions, the TCC on the left in Figure 1 may be considered to provide an estimate of ζ.

To continue with this illustration, let's say the TCC on the right in Figure 1 is based on a subset of items non-randomly selected from the item bank for the operational test. For example, the items may have been chosen in part because they helped target the test around the cut score. It should be noted that the TCC of the operational test is generally more difficult than the item bank. This is indicated by the fact that the TCC of the operational test is displaced to the right of the TCC for the item bank.

In Figure 1 it should be clear that the observed domain score of $Z=.70$ (based on the operational test) is no longer an unbiased estimate of ζ. However, an unbiased estimate of ζ can still be obtained by referencing Z to the item bank TCC. The justification for this is that all items in the bank have been calibrated on a common scale. Since the operational test is a subset of these items, the θ obtained on the operational test can therefore be mapped on to the item bank domain scale via the item bank TCC. In this way an estimate is obtained of what the examinees' true score would have been had s/he answered all items in the bank.

The way this is accomplished in Figure 1 is that the operational TCC is first used to transform the examinees' observed Z of .70 to a θ=1.5. Then, the item bank TCC is used to map the examinees' θ of 1.5 onto the item bank domain scale. We see that even though the observed domain score is Z=.70, the estimated true domain score is ζ=.85. The discrepancy between the observed and estimated true domain scores is due to the greater difficulty of the operational test.

Another use of this approach is possible when a test is divided into subtests and it is reasonable to assume that each measures the same trait. Under these conditions each subtest domain score can be placed on a common underlying θ scale and referenced to a single item bank TCC. This facilitates criterion-referenced comparisons between subtests and makes more meaningful the relationship between subtest and total test domain scales.

When all items on a test have been calibrated on a common θ scale than each subtest provides an independent estimate of the examinees' θ. However, each subtest domain score may vary because of differences in subtest difficulty (in the two- and three-parameter models we also may assume subtest differences in degree of discrimination and guessing). This causes special problems in comparing subtest performance. As an example, an examinee's estimated true score of .70 in domain A may mean

more ability than an estimated true score of .80 in domain B if the domain A subtest is more difficult than the domain B subtest.

The only way to compare performance on the subtests is to reference them to a common TCC. Probably the most convenient TCC is the overall item bank TCC which is based on all items in the item bank. By referencing all subtest observed domain scores to the overall item bank TCC diagnostic information can be obtained about the examinees' strengths and weaknesses in each subtest relative to his/her overall ability.

Unidimensionality Assumption

One of the assumptions underlying the above procedures is that each subtest measures the same trait. Currently, there is no satisfactory way of testing this assumption within the context of latent trait models (except for the fact that model-data fit is itself a test of this assumption). However, within the context of classical test theory this assumption can be tested by fitting a LISREL congeneric measurement model to the subtest variance-covariance matrix (Jöreskog, 1971). The following is a proposal of how this might be done. To keep the discussion consistent with LISREL notation the observed subtest raw score X rather than the domain score Z will be used. The true subtest raw score will be denoted by ξ.

Within Jöreskog's general LISREL model the matrix of q observed subtest domain scales, S_X is assumed to have the following variance-covariance structure.

$$S_x=\Lambda_x\Phi\Lambda_x'+\Theta_\delta \ .$$

In the above equation Λ_X is a q×n factor pattern matrix (where q=the number of subtests, and n=the number of factors), Φ is an n×n variance-covariance matrix of factors, and Θ_δ is a q×q diagonal matrix of measurement error variances. The linear model underlying the above observed variance-covariance structure is

$$X=\Lambda_x\xi+\delta \ .$$

In the linear model X is a q×1 vector of observed subtest variables. ξ is an n×1 vector of subtest factors, and δ is a q×1 vector of measurement error variables. As an example of the linear model, if a criterion-referenced test has five subtests, then the elements of the linear matrix equation are

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix}$$

The above matrix equation specifies five subtests, X_1, X_2, .., X_5, each measuring an underlying trait, $\xi_1, \xi_2, .., \xi_5$.

A congeneric measurement model can be fitted to the data by specifying $\Phi=I$ (where I equals an identify matrix). This specification states that the true domain scores underlying each subtest are perfectly correlated and is identical to specifying that there are only n=1 factors underlying all five domains, i.e., $\xi_1=\xi_2=\xi_3=\xi_4=\xi_5$. If one wishes to test more restrictive classical test theory measurement models this can also be done with LISREL. For example, a tau-equivalent model may be tested by assuming $\lambda_1=\lambda_2=\lambda_3=\lambda_4=\lambda_5$. In addition, a parallel measurement model may be tested by further assuming $\delta_1=\delta_2=\delta_3=\delta_4=\delta_5$.

Although good fit of a congeneric measurement model provides additional evidence of unidimensionality, it is not a strict test of the unidimensionality of latent trait models. The reason for this is because classical test theory assumes the regression of observed scores on true scores is linear, whereas, latent trait models assumes the regression is curvilinear.

SUMMARY

This paper presented an introduction to the use of latent trait models for the estimation of domain scores. It was shown that these models provide an advantage over classical test theory and binomial error models in that unbiased estimates of true domain scores can be obtained even when items have not been randomly selected from a universe of items. This allows the full use of item statistics during test development. Furthermore, it was argued that when all domain subtests measure the same underlying trait, then all subtest domain scales may be related to each other via the underlying θ scale.

Finally, a method was presented whereby the unidimensionality assumption can be tentatively tested. This was to fit a LISREL congeneric measurement model to the observed subtest variance-covariance matrix.

REFERENCES

Brennan, R. L. "Applications of Generalizability Theory." In Ronald A. Berk (ed.), Criterion-Referenced Measurement: State of the Art. Baltimore, Maryland: The Johns Hopkins University Press, 1980.

Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. B. "Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments." Review of Educational Research, 1978, 48, 1-47.

Huynh, H. "On the Reliability of Decisions in Domain Referenced Testing." Journal of Educational Measurement, 1976, 13, 253-264.

Jöreskog, K. G. "Statistical Analysis of Sets of Congeneric Tests." Psychometrika, 1971, 36, 109-133.

Lewis, Charles; Wang, Ming-Mei; and Novick, Melvin R. "Marginal Distributions for the Estimation of Proportions in M Groups." Technical Bulletin, No. 13. Iowa City: American College Testing Program, 1973.

Livingston, S. A. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Lawrence Erlbaum, 1980.

Millman, Mason. "Criterion-Referenced Measurement." In W. James Popham (ed.) Evaluation in Education: Current Applications. Berkeley, California: McCutchan, 1974.

Millman, Jason. "Determining Test Length: Passing Scores and Test Lengths for Objectives-Based Tests." Los Angeles: Instructional Objectives Exchange, 1972.

Novick, Melvin R.; Lewis, Charles; and Jackson, Paul H. "The Estimation of Proportions in M Groups." Psychometrika, 1973, 38, 19-46.

Novick, M. R. and Lewis, C. "Prescribing Test Length for Criterion-Referenced Measurement." In C. W. Harris, M. C. Alkin, and W. J. Popham (eds.) Problems in Criterion-Referenced Measurement (CSE monograph series in evaluation, No. 3). Los Angeles: Center for the Study of Evaluation, University of California, 1974.

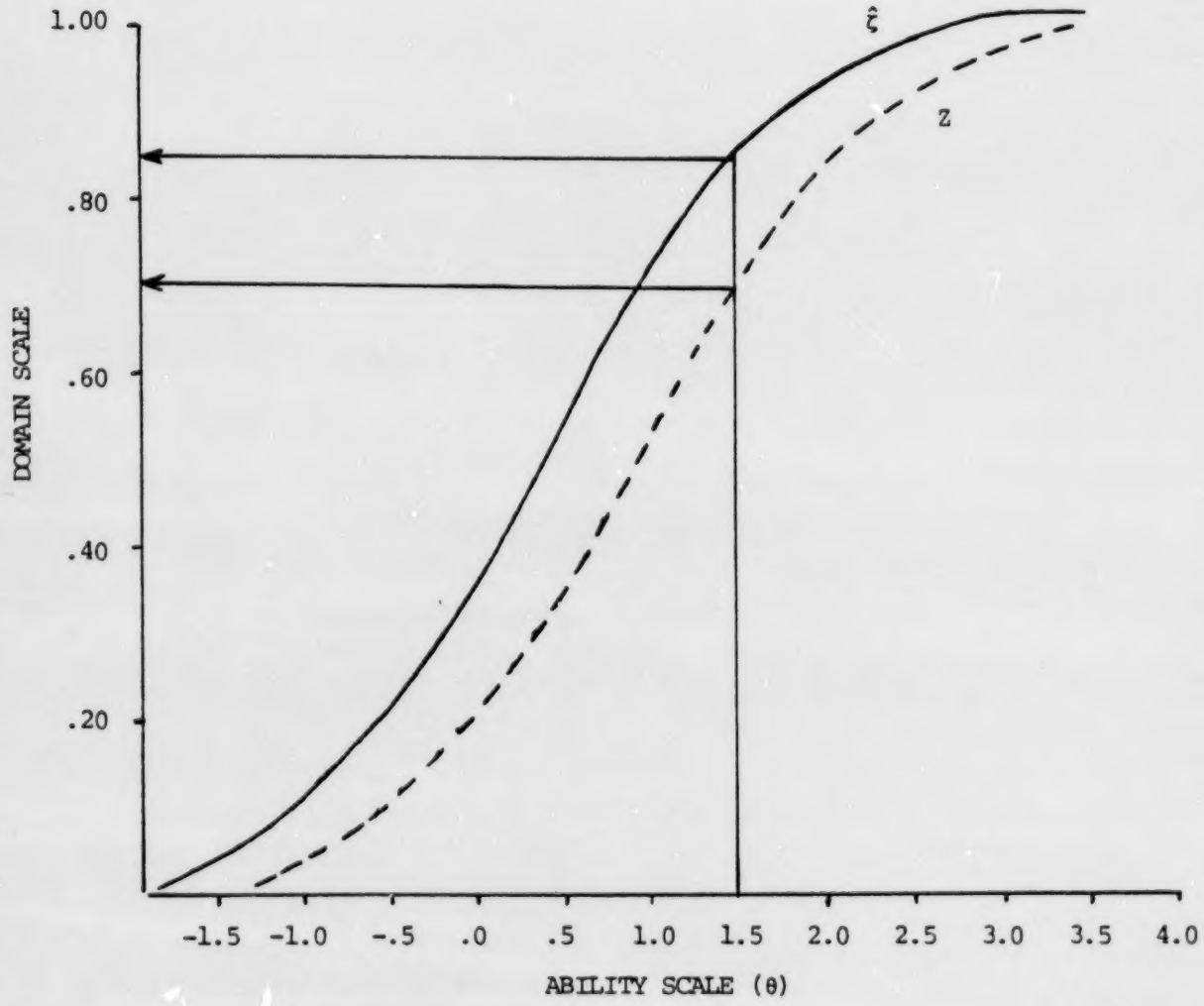

Figure 1. Test characteristic curves for estimated true scores $\hat{\zeta}$ and observed domain scores Z resulting from an item bank and operational test, respectively.